Capital Bike Share Predictive Model Report

Prepared By: Clark P. Necciai Jr.

November 17, 2023

Contents

1 Executive Summary

Capital Bike Share provides a network of multi-purpose bicycles to the denizens of the Washington D.C. Metropolitan region. We were approached by Capital Bike Share to delve into their hour-by-hour observations across 2011 and 2012. Preliminary insights into the dataset revealed that demand usage for these bicycles can be affected by a variety of influential factors. We determined two major categories of influence: time and weather.

We examined multi-variable trends and deliberated on insightful patterns. Based on our findings, we concluded rental demand as being mainly determined by registered riders using bicycles as a primary mode of transportation to and from work. Furthermore, our resulting predictive model confirmed our preliminary analysis, as it was ultimately determined that time-based variables were the most significant factors in accurately predicting bicycle rentals.

2 Problem Statement and Approach

Our primary tasks in this analysis were twofold:

- Identifying the most influential variables relative to their predictive power in determining the total hour-by-hour bicycle rentals
- Fitting a multiple regression model predicting total hour-by-hour bicycle rental count

Beginning with an exploratory data analysis, we aimed to determine variables which we believed have significant relation to the target variable. After a well-documented, granular analysis of 17 features over 17,379 observations, we utilized our findings for comprehensive modeling. This process culminated in a multiple linear regression model that not only accurately predicted rental demand, but simultaneously provided an assumption-backed evaluation confirming our models' reliability and generalizability. We then end with our conclusions and recommendations.

3 Methodology

3.1 Data Preprocessing

We began our approach with an overview of the integrity of our dataset's structure. We discovered no missing values nor duplicated observations. We did, however, note variables which had data types that were non-representative of their underlying values.

Variables `dteday`, `season`, `yr`, `mnth`, `hr`, `holiday`, `weekday`, `workingday`, and `weathersit` were all considered to have data types and values which were non-representative. As such, we applied more representative data types, effectively categorizing them and applying appropriate labels. The `dteday`, `yr`, `mnth`, `hr`, `weekday`, `weathersit`, and `cnt` were renamed for additional clarity.

3.2 Exploratory Data Analysis (EDA)

To better understand our variables, we examined each of our variables' summary statistics and distributions while investigating for patterns, inconsistencies, and anomalies which may affect our analysis.

3.2.1 Target Variable

3.2.1.1 `count_rentals` The variable we aim predict is `count_rentals`(`cnt`). Each observation is taken on an hourly basis. Our target variable's distribution shows us that `count_rentals` is highly right skewed. There is serious variability in the hour-by-hour rentals, with a minimum of a single rental to a maximum of nearly a thousand rentals per hour, but with 50% of our observations being less than the median of **142** bicycle rentals per hour, marked by the red dashed line.

3.2.2 Predictor Variables

Each categorical, continuous, and Boolean features' summary statistics and distributions were thoroughly investigated. Categorical features we considered are as follows: `season`, `year`, `month`, `day`, `hour`, `date`, and `weather`. Continuous features are `temp`, `atemp`, `hum`, and `windspeed`. Boolean features were `holiday` and `workingday`.

The distributions and corresponding summary statistics for the categorical, continuous, and Boolean variables can be observed in Figure X, Y, and Z respectively.

3.2.2.1 Noteworthy Observations

3.2.2.1.1 `date` The `date` distribution revealed that not every unique value of date has an expected equivalent number of hourly interval measurements. The majority of each of the **731** distinct date values contained twenty-four or twenty-three of the expected hourly observations. However, fourteen of the dates contained fewer, which in turn equated to **103** hours worth of missing possible observations. Capital Bike Shares' hour-to-hour observational system has hourly gaps, which would otherwise provide useful statistics.

3.2.2.1.2 `holiday` Inspection of our `holiday` variable revealed dates which should have been marked as holidays and others which should not have been. We re-labeled these observations based on official federally recognized holidays.

In comparison to our dataset: 2011-01-01, 2011-12-25, 2012-01-01, and 2012-11-11 were incorrectly mislabeled as *not* being holidays. 2011-12-26, 2012-01-02, and 2012-11-12 were mislabeled as **being** holidays. 2011-04-15 and 2012-04-16 are dates of observance for Emancipation Day in the Washington D.C. Area. Due to the holiday-like observance of Emancipation Day in our area of interest, these two dates will be labeled as holidays.

3.3 Correlation

We decided to investigate the extent to which our variables are linearly related to one another with a visualized correlation matrix[Figure XX]. Here we are concerned with those relationships related to that of our target, `count_rentals` and continuous predictor variables. Here we list the most noteworthy findings:

- Being that `temp` and `atemp` both aim to measure normalized Celsius, it is unsurprising that these variables exhibit a near exact linear relationship with one another. We additionally see they both are moderately, positively related to our target variable to the same degree. Naively including both `temp` and `atemp` in our modeling process would result in multicollinearity. To maintain a robust regression model and avoid multicollinearity which would cause our coefficients estimates to become unstable, we will choose only one of these variables, `atemp`, to be included in the modeling process.
- Our target is slightly negatively correlated with humidity. Being that humidity includes potential precipitation such as rain, mist, snow, and others, this relationship is unsurprising.

Overall, we see that each of our continuous predictor variables exhibits straight-line relationship with our target variable. This can be seen by examining the bottom row of Figure XX as indicated by scatter-plots and the red trend lines.

3.4 Notable Multivariate Analysis

3.4.1 Temperature and Target Relationship

We examined the multi-variable relationship between our temperature related variables, `temp`, `atemp`, and `count_rentals`. We revealed twenty-four high-leverage observations having the exact same `atemp` value of 0.2424, but `temp` values which were relatively high and variable in relation, ranging from [0.62 to 0.86]. When considering the observational evidence of these twenty-four `temp` and `atemp` observations as being high leverage outliers, we have opted to remove these observations. This removal helps to ensure data integrity as we move towards the modeling process.

3.4.2 Work Commute - Hourly Rental Trends

Our intuition led us to investigate possible patterns of usage based solely on time. What we uncovered, showed that during working days, there is strong indication that bicycles are being utilized as a primary mode of transportation to and from work. Most notably, it is Capital Bike Shares `registered` rental users that are driving this trend. We see that between all rentals and registered rentals, the trend is nearly identical[Figure XX, Figure XX]. Neither `casual` rentals nor rentals occurring on Saturday/Sunday exhibit this trend, adding further evidence to our speculation[Figure XX, Figure XX]. The bicycles may be being used as a primary mode of transportation for the majority of `registered` riders during to/from work. Around 8:00AM we note an upward trend of rentals and a decrease around 5:00PM/6:00PM.

3.5 Feature Reduction

`date` - Our team reasoned that the creation of **731** distinct variables representing `date` could have severe impacts (high VIF and overfitting) on model performance. No unique information is provided by `date` that is not already found in our `year` and `month`, and `day` variable. To avoid these drawbacks, we have decided to disregard the `date` variable from the modeling process.

Additionally, being that `instant` is merely an identifier for the observations, it will also be disregarded during the modeling process. Lastly, being that `count_rentals` is the exact sum of `casual` and `registered`, inclusion of these two would make our prediction analysis arbitrary. We likewise disregarded these from our modeling.

3.6 Feature Engineering

Dummy Variable Creation - Categorical variables, such as `season`, `year`, `month`, `day`, `hour`, and `weather`, are better suited for dummy variable creation for use in our multiple linear regression model. This encoding ensures these factor/categorical variables are appropriately interpreted by our model.

4 Model Building

4.1 Data Partition

With our goals centered on finding the most significant predictors with respect to determining the total number of hour-by-hour rentals, we decided to partition our data 80% towards training the model and 20% for testing model performance. Allowing a significant majority of our overall dataset to go towards the model fitting process enables us to get more precise, stable coefficient estimates.

4.2 Initial Full Model(s)

Our initial, comprehensive full model encompasses all possible predictors within the Capital Bike Share dataset. Our approach allows us to gain a generalized sense as to the importance and effects our predictors are having in determining our response variable.

Initial Full Model

$$\hat{count rentals} = \beta_0 + \beta_1\texttt{holiday} + \beta_2\texttt{workingday} + \beta_3\texttt{atemp} + \ldots + \beta_{52}\texttt{weather\_Type\_4}$$

Where $\hat{count rentals}$ is the predicted number of rentals, β_0 is the intercept, and $\beta_1, \beta_2, ...\beta_n$ are the coefficients of the predictor variables.

While the full model containing all possible predictors can provide some insightful information, it is far from refined. Many of the predictors were found to be statistically insignificant in predicting the response. While we could remove these variables, we will allow

step-wise selection to determine the best set of predictors in determining the response. To refine this full model, we will first inspect the assumptions of multiple linear regression.

4.2.1 Model Diagnostics

4.2.1.1 Normality of Residuals Our residuals should be normally distributed and can be visualized using a Q-Q plot. Deviations from our straight line in the diagnostic plot would suggest potential non-normality and is confirmed by the formal Kolmogorov-Smirnov Test. We can be certain that our residuals are not normally distributed.

4.2.1.2 Homoscedasticity The homoscedasticity assumption states that we should have residuals (ϵ) with a constant variance. The funnel-shape we see in our diagnostic plot is indicative of the opposite, heteroscedasticity, or non-constant variance, and is confirmed by our formal Breusch-Pagan Test. Our standard errors, confidence intervals, and subsequently our hypothesis testings rely on the homoscedasticity assumption.

4.2.1.3 Linearity The linearity assumption states that we should assume the true relationship between our predictors and response variable is a straight-line. We can identify non-linear trends with the red line fit to our residuals. Linearity appears to be violated here, as the upward-curved line is indicative of a non-linear relationship.

4.2.1.4 Multicollinearity The presense of multicollinearity reduces the accuracy in our model's coefficients by causing our coefficients' standard errors to grow, effectively masking their importance and making interpretation difficult. We can detect multicollinearity by calculating the variance inflation factor of our model's predictors. Our first full model contains a few predictors having VIF values greater than 10 which warrants possible removal from our model to remove multicollinearity.

4.2.1.5 Independence of Errors Our residuals should be independent of each other for our regression to be reliable. We can see from the previously mentioned Residuals vs. Fitted Figure that we have evidence of having violated this assumption and can confirm that indeed we have a lack independence of errors from the Durbin-Watson Formal Test. Our test result shows that indeed there is strong evidence of positive autocorrelation.

4.2.2 Model Improvement

We want our final multiple linear regression model to be robust. To achieve this and bring our assumptions closer to be satisfied, we now proceed with various approaches of addressing issues which may be affecting our model's assumptions.

4.2.2.1 Remove Influential Observations We begin by removing observations which we believe may be having too disproportionate an impact on our model's fit. The observations in particularly are likely to affect our model's coefficients, affecting our overall predictive accuracy. These include the three observations with significantly high cook's distances. Each

removed observation were of `weather_type_4`. Because of this, the `weather_type_4` variable was consequently removed from the training data set.

4.2.2.2 Target Variable Transformation We considered multiple transformations to approach normality and address the linearity and homoscedasticity assumptions, including, the Box-cox, logarithmic, square-root, cube-root, and fourth-root transformations. For each of our target variables, the transformation which approximated normality (skewness value to 0) the closest was the **cube-root** transformation. Therefore, we will predict the cube-root transformed version of our target variable `cube_root_total` for the final model.

4.2.2.3 Multicollinearity We opted to remove the influences of multicollinearity in our models via removing the variables which had the highest variance inflation factors(VIF > 10) one at a time. The two variables that were removed were ultimately removed were `workingday`, and `season_Summer`.

4.3 Final Reduced Model Using Stepwise Selection

Using step-wise selection, we identified an optimal subset of predictors. These variables were determined to be the best set of predictors in predicting the transformed response, `cube_root_total`.

Our team believes that two of the best metrics for determining model fit are Adjusted R^2 and Bayesian Information Criterion (BIC). Both of these metrics help determine models which balance goodness-of-fit and complexity. We seek to maximize Adjusted R^2 and minimize BIC. Our step-wise selection found a final model which meets that criterion and can be seen in the Appendix as Figure Z with a visualization depicting the change in R^2 and BIC in Figure Z.

4.3.1 Interpretation of Model Coefficients

From our step-wise model, the three variables with the most significant impact on the expected number of total rentals are `hour_5:00PM`, `hour_6:00PM`, and `hour_8:00AM`.

For instance, when our `hour_5:00PM` variable is 1(True), holding other variables constant, the cube root number of bicycle rentals is expected to increase by 4.22167. In other words, if we back-transform(apply cubic) our coefficient value for interpretability, the number of bicycle rentals is expected to increase by 75.24. This procedure of coefficient interpretation applies to all coefficients in our model. In the case of continuous variables, such as `atemp`, the expected increase in the number of bicycle rentals is determined by a one unit increase in the predictor value.

4.3.2 Model Diagnostics

Our approaches to satisfy the assumptions mentioned during the first full model fitting were promptly addressed with our transformation of the response variable via a cube root transformation, removal of over-influential observations, and high variance inflation variables. The

Despite our approach, we did observe some potential violations of regression assumptions in homoscedasticity, linearity, and normality/independence of residuals. The multicollinearity assumption was satisfied due to our removing of high variance inflation variables.

In practice, however, the assumptions are rarely validated. After employing our corrective measures and observing the resulting vastly improved diagnostic plots, we remain confident in our model's predictive capabilities and generalizability. Our attempts at rectifying our assumptions helped to stabilize our coefficients, and ensure that our model contains variables meaningful in determining our target.

4.3.3 Test Set Evaluation - Model Performance

Utilizing the final step-wise model determined above, we tested our model's performance by making predictions on the test set. Because we applied a cube-root transformation to our target variable `cube_root_total`, we **back-transform** these predicted/fitted values for interpretability. Doing so allows us to draw meaningful interpretation from our model's performance in the original units.

We have provided the Root Mean Squared Error(RMSE), R^2, and Mean Absolute Error(MAE). RMSE and MAE are measured in the same units as the target variable and thus have a meaningful interpretation. MAE for example, is the "on-average" error between the predicted value and the true number of bike rentals, which for our model is approximately **63** rentals. Additionally, our R^2 value is approximately 0.734, meaning that 73.4% of the variability in the number of bike rentals is explained by our model which tells us that our model is a strong fit and is generalizable to new data predictions.

5 Conclusions

The Capital Bike Share dataset analysis revealed insightful patterns and statistics in pursuit of variables which were used in formulating a reliable multiple linear regression model. Significant influences found when examining the distributions of features such as the hour of the day and real feel Celsius temperature, help us understand what drives the number of total rentals.

Our final model provides a concise selection of key predictors, allowing us to filter out non-meaningful features which would otherwise prove irrelevant. Our approach allows us to ultimately make accurate, and reliable predictions as to the number of Capital Bike Share rentals.

5.1 Recommendations

Based on the hourly working day commute trends showing commutes to and from the workplace as being a major driving force in the number of rentals, and the extreme prevalence and significance of the hour variables present in our final model, we believe we have a strong recommendation for Capital Bike Share.

There can be no doubt that registered commuters are utilizing bicycles as their primary mode of transportation during workplace commutes. Capital Bike Share can capitalize on this, by installing/placing additional bicycle kiosks in key residential and traffic heavy work-place zones. Doing so would incentivize additional bike rentals by permitting commuters a reliable source of transportation and ultimately increasing the number of bicycle rentals.

6 Appendix

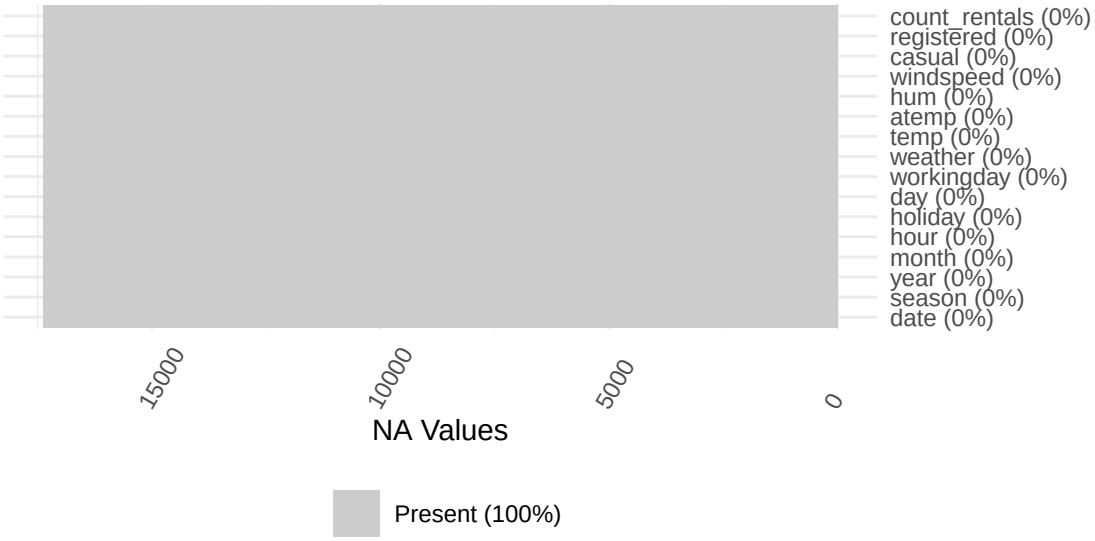

Figure 1: No missing values in dataset

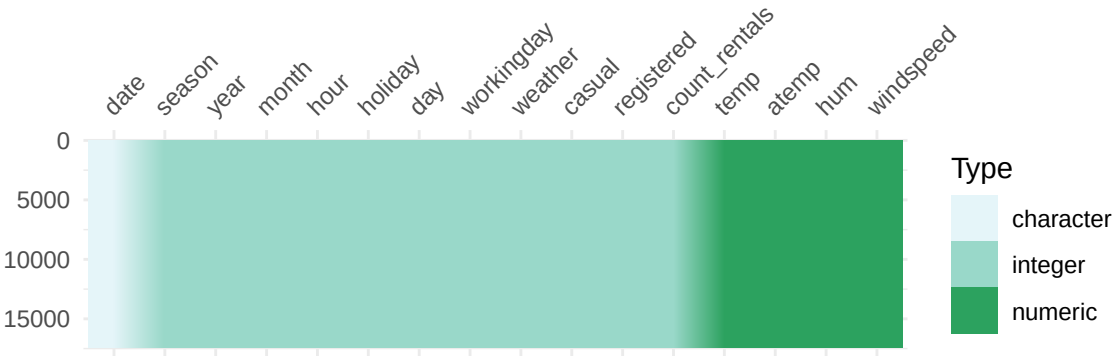

Figure 2: Pre-assigned Variable Data Types

\begin{figure}

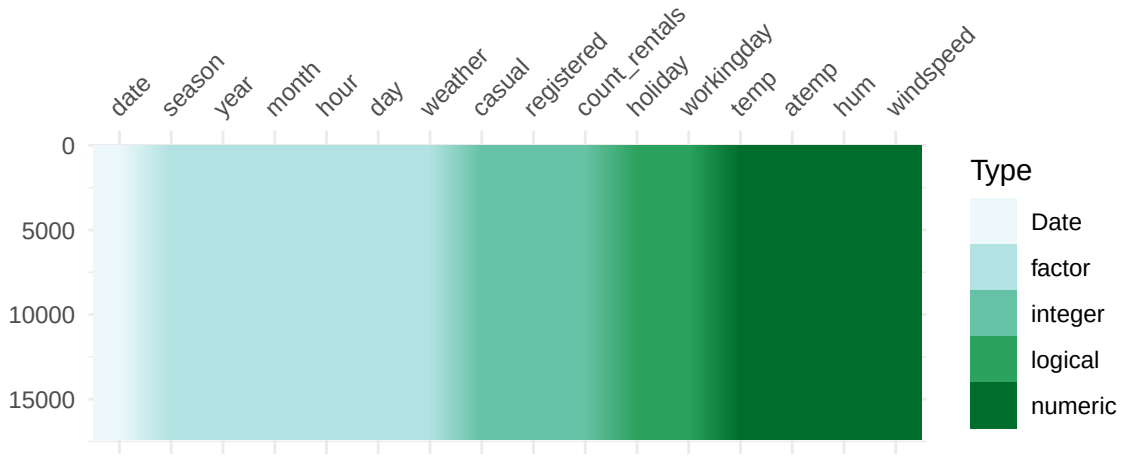

Figure 3: Reassigned Variable Data Types

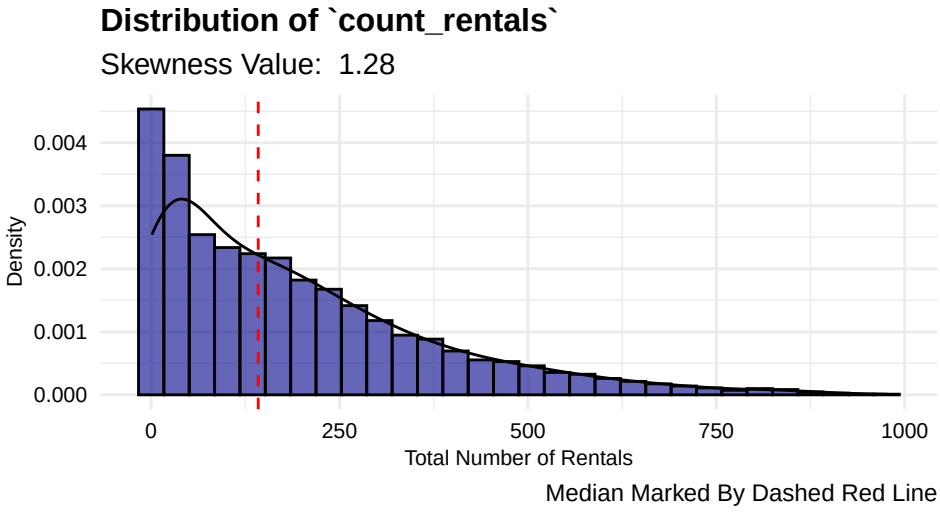

Figure 4: Distribution of Target Variable

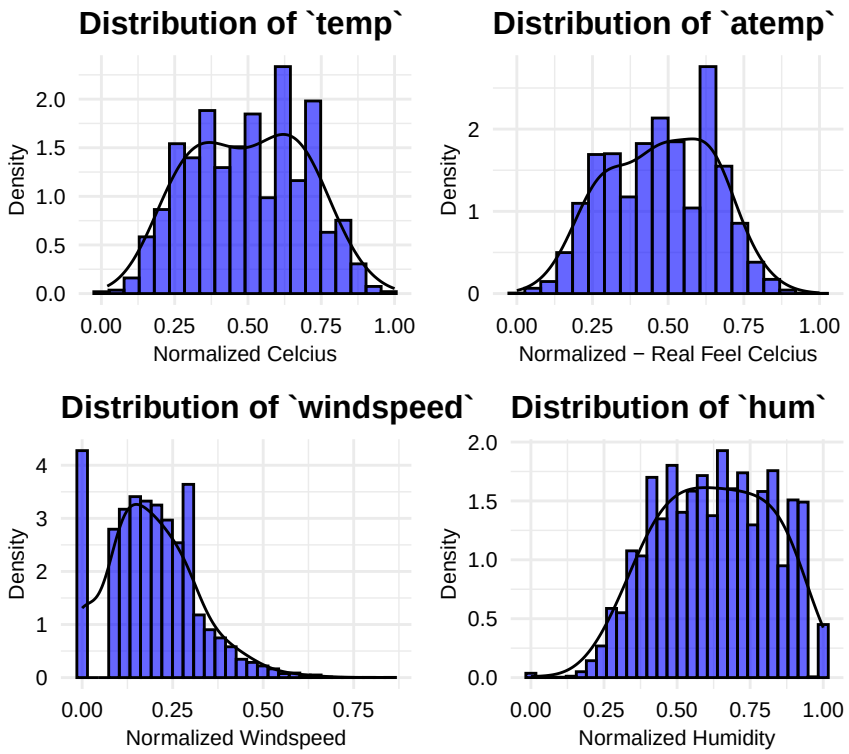

Figure 5: Continuous Variables Distribution

	Mean	Median	StdErr	Skew	Q1	Q3	IQR	Min	Max
count_rentals	189.46	142	181.39	1.28	40	281	241	1	977
temp	0.5	0.5	0.19	-0.01	0.34	0.66	0.32	0.02	1
atemp	0.48	0.4848	0.17	-0.09	0.3333	0.6212	0.2879	0	1
hum	0.63	0.63	0.19	-0.11	0.48	0.78	0.3	0	1
windspeed	0.19	0.194	0.12	0.57	0.1045	0.2537	0.1492	0	0.8507

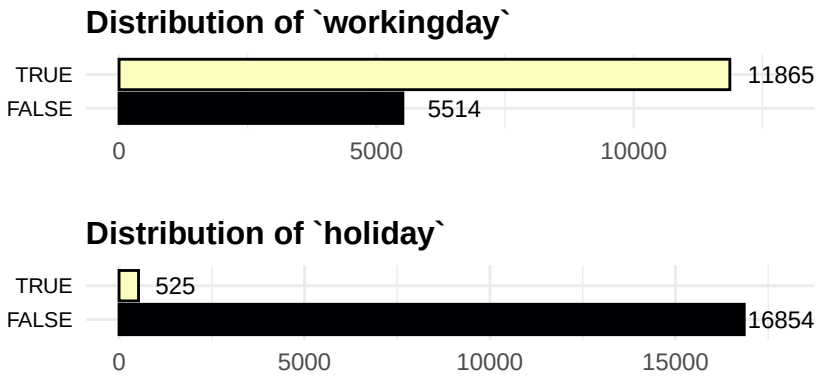

Figure 6: Boolean Variables Distribution

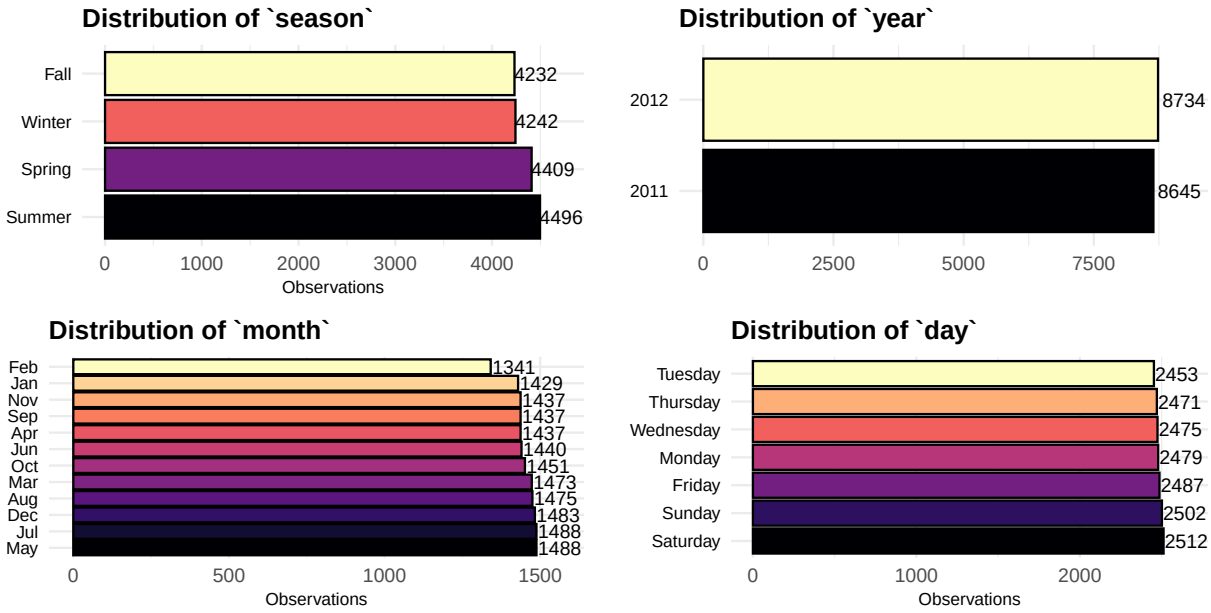

Figure 7: Distributions of ‘season‘, ‘year‘, ‘month‘, and ‘day‘

```
## NULL
```

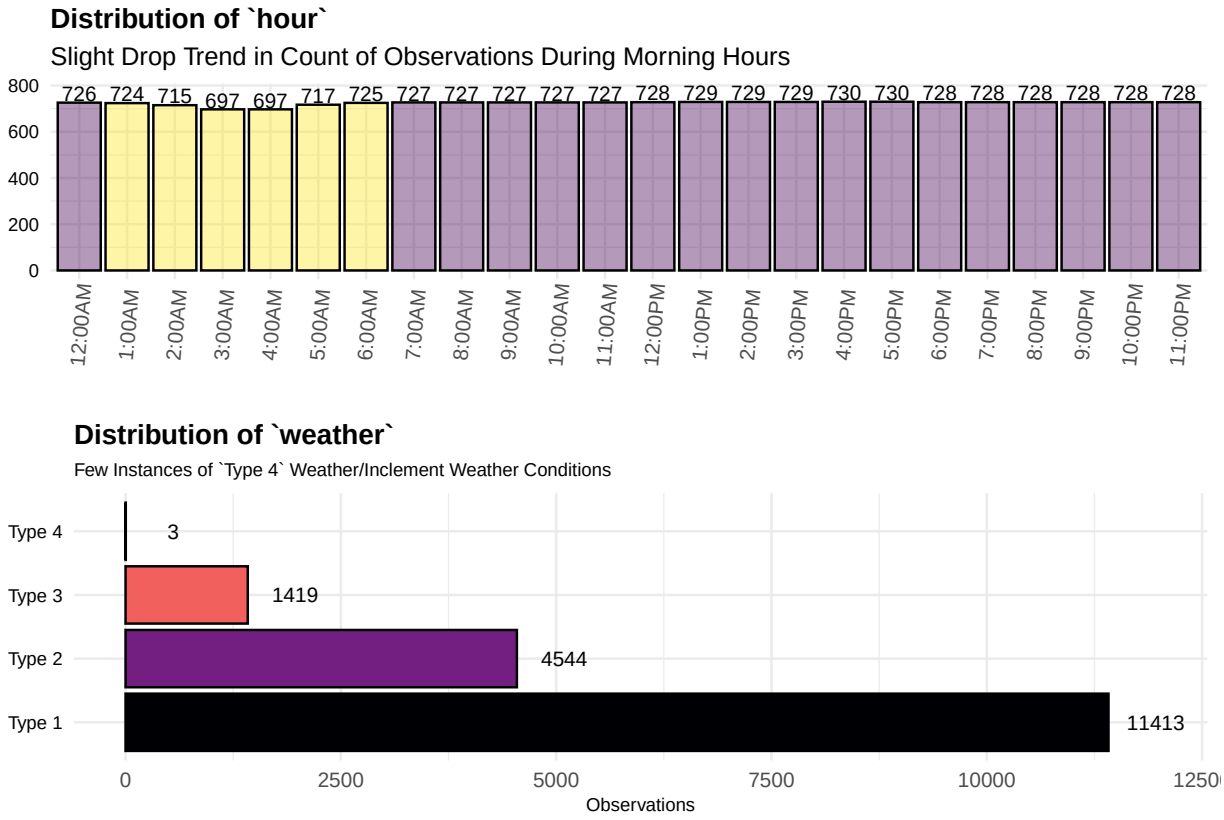

Figure 8: Distributions of 'hour', and 'weather'

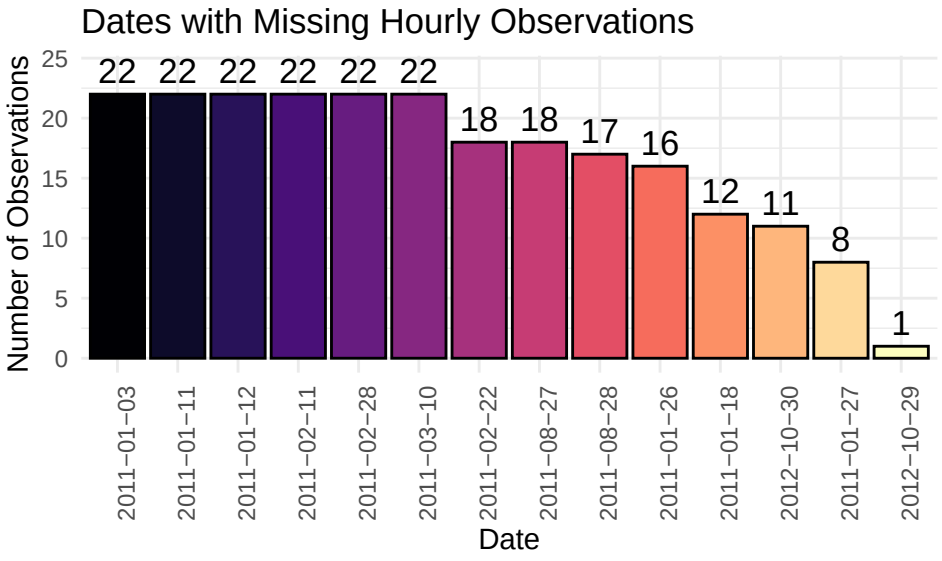

Figure 9: Dates with Low Number of Observations

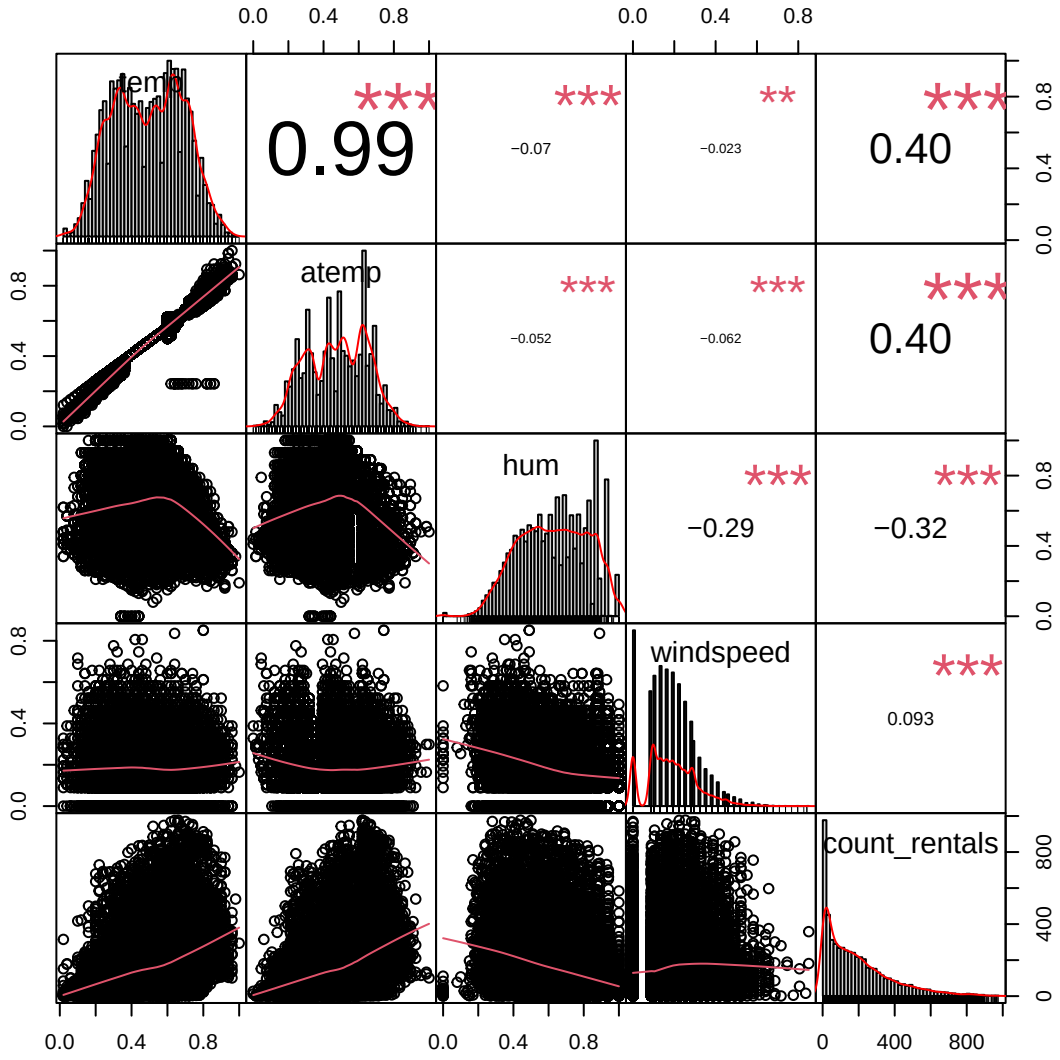

Figure 10: Correlation Matrix and Continuous Variables Relationship with Target Variable

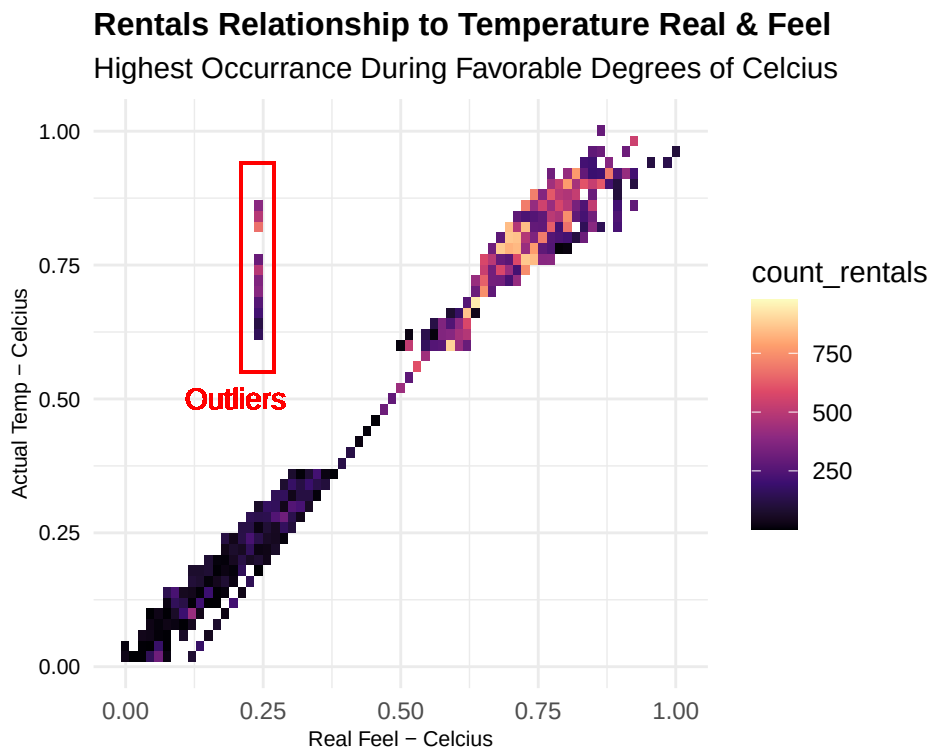

Figure 11: High Leverage Points Between ‘temp‘ and ‘atemp‘

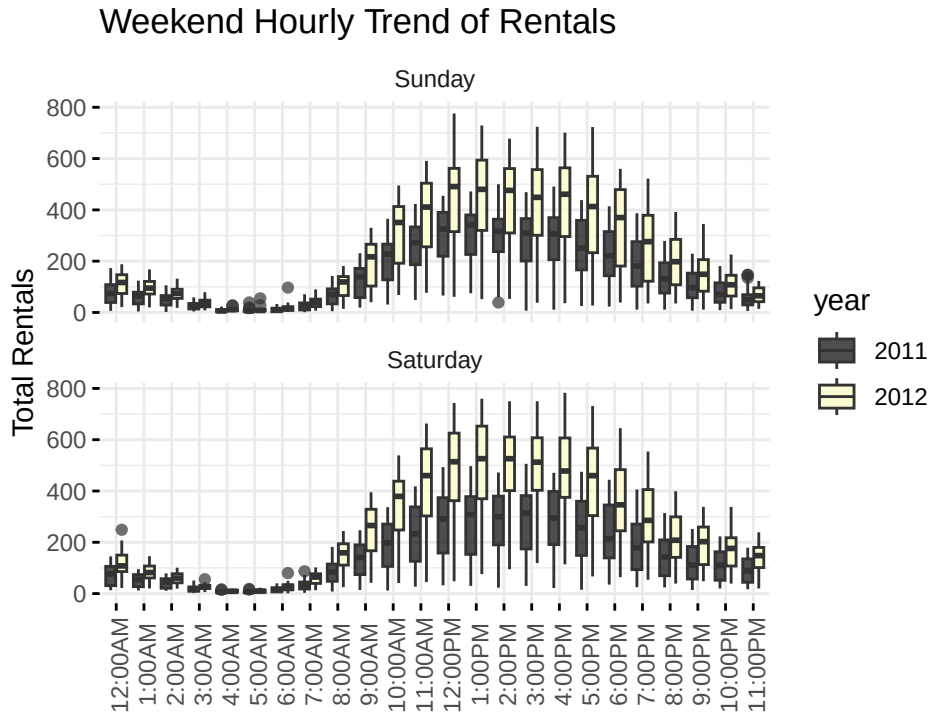

Figure 12: Smooth Curve of Rental with Peak Usage Around Midday

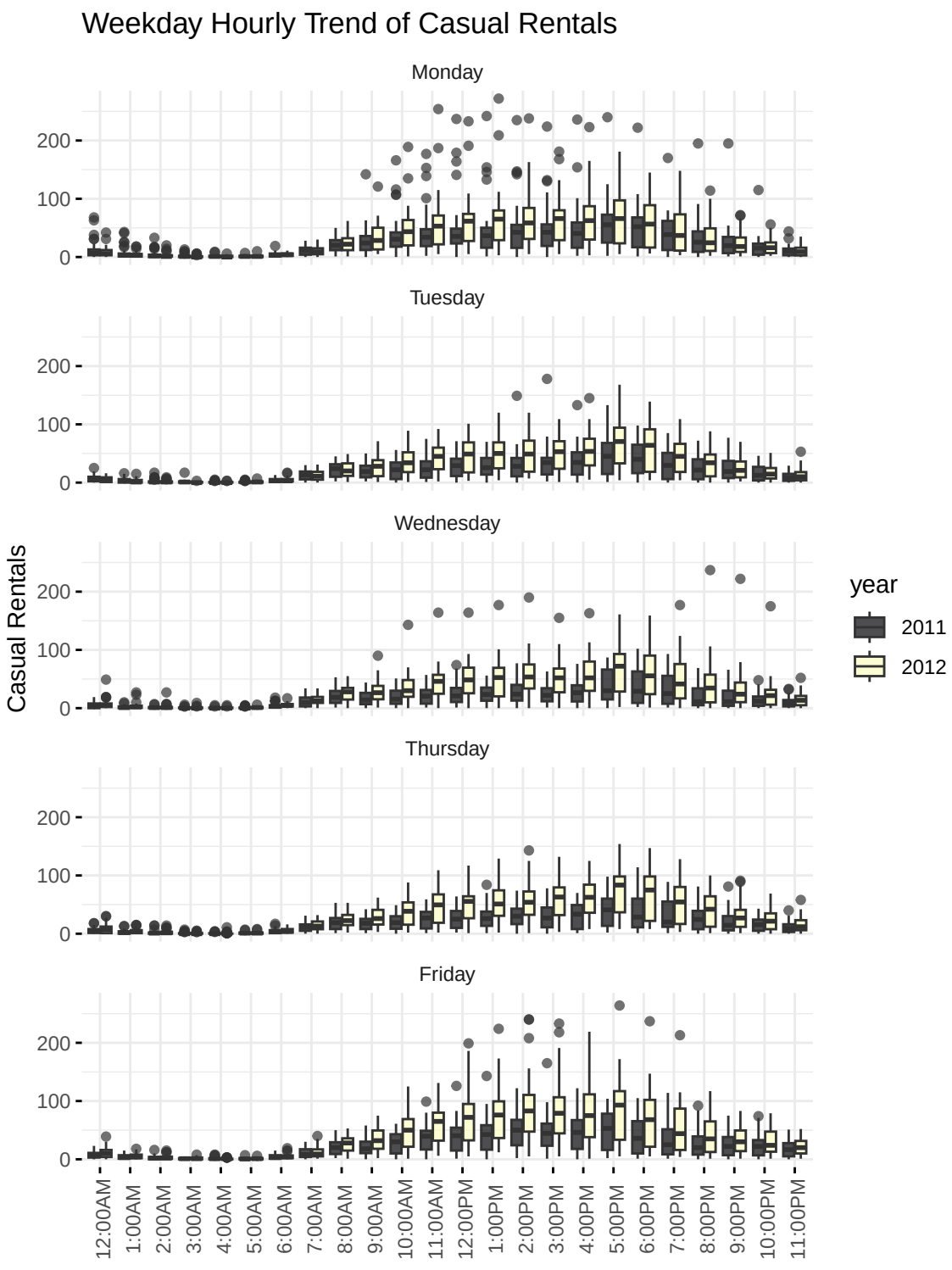

Figure 13: Low Casual Usage Throughout Working Days

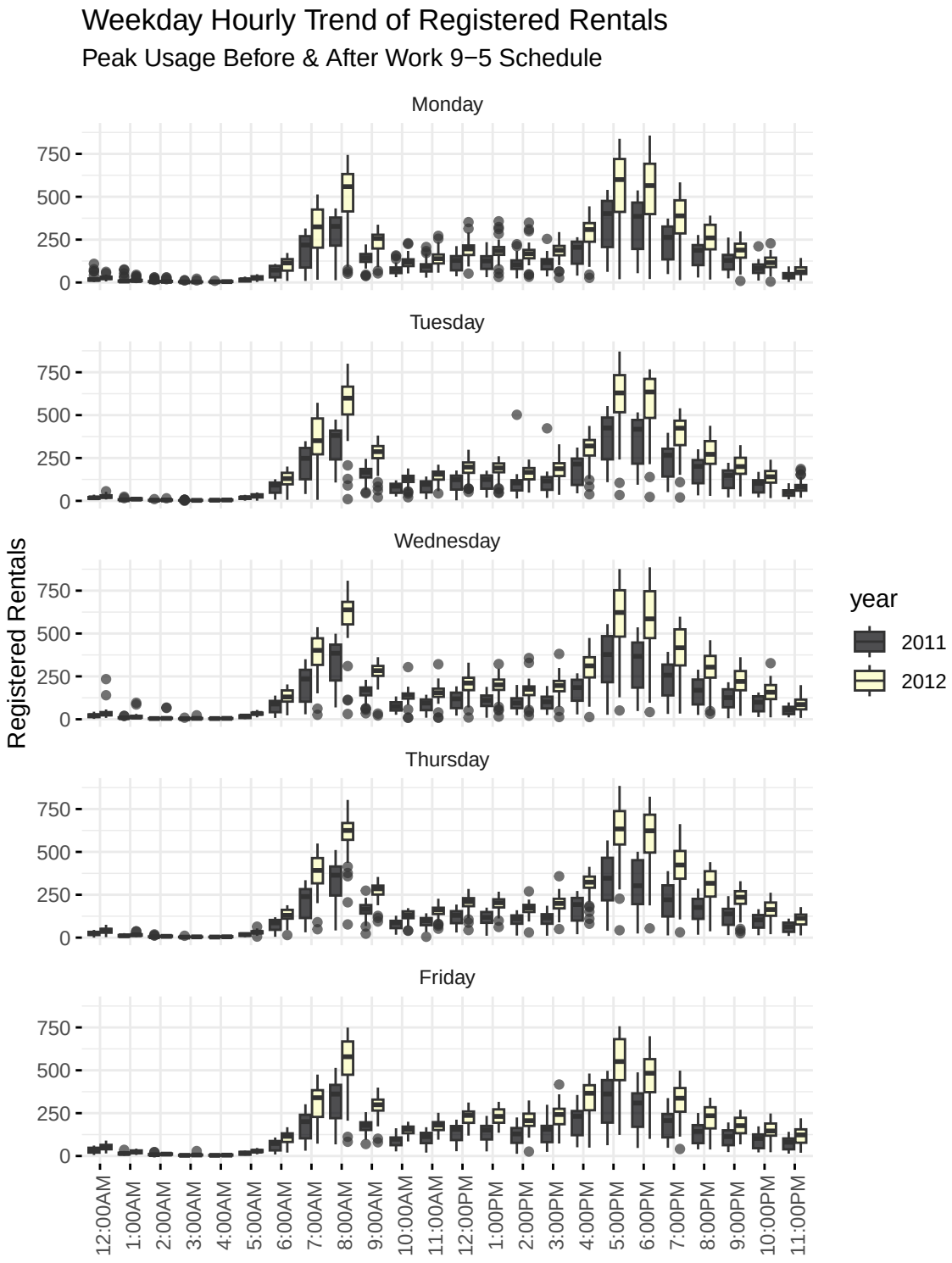

Figure 14: Registered Work Commute Usage

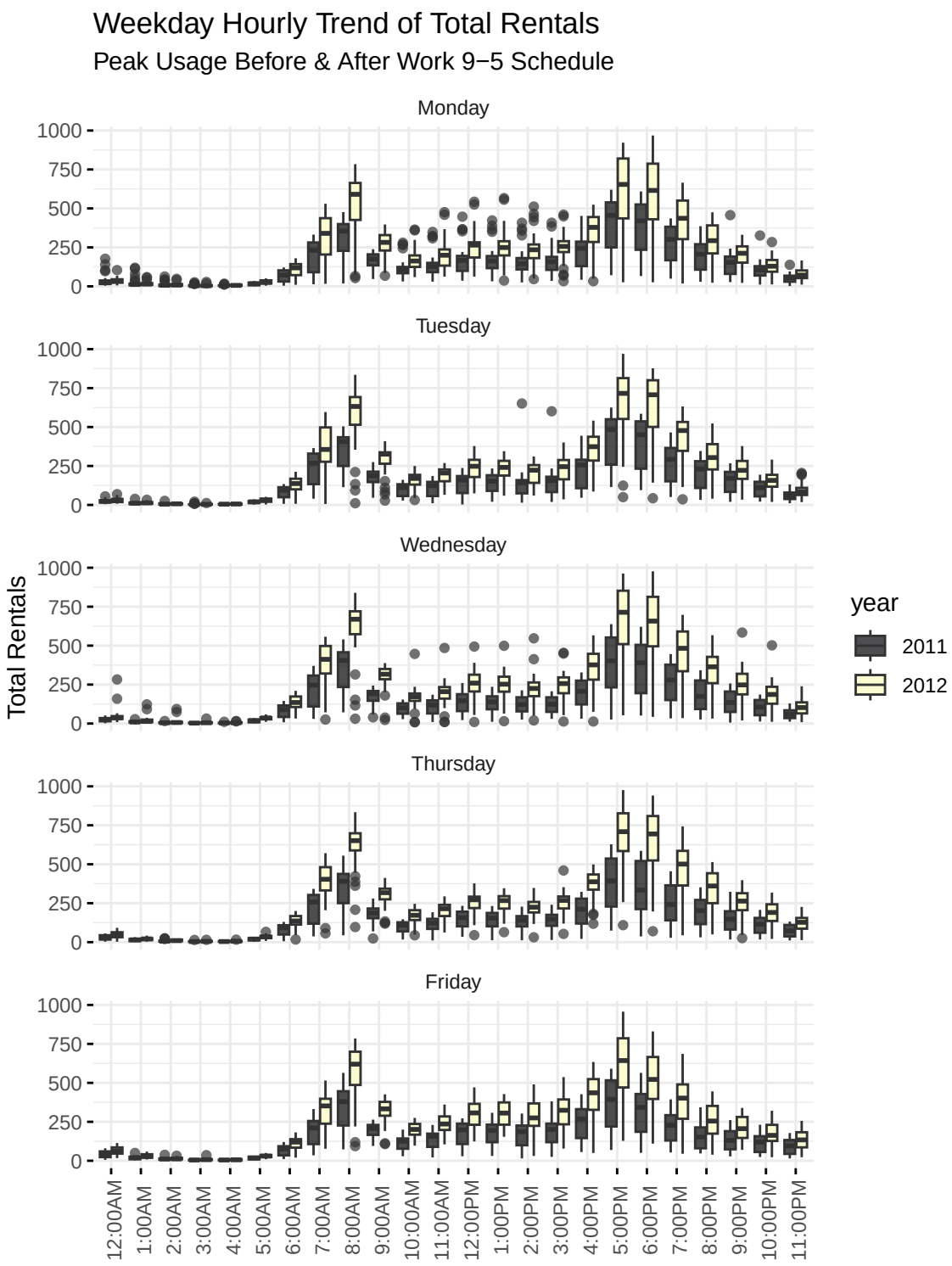

Figure 15: Total Rentals Work Commute Usage

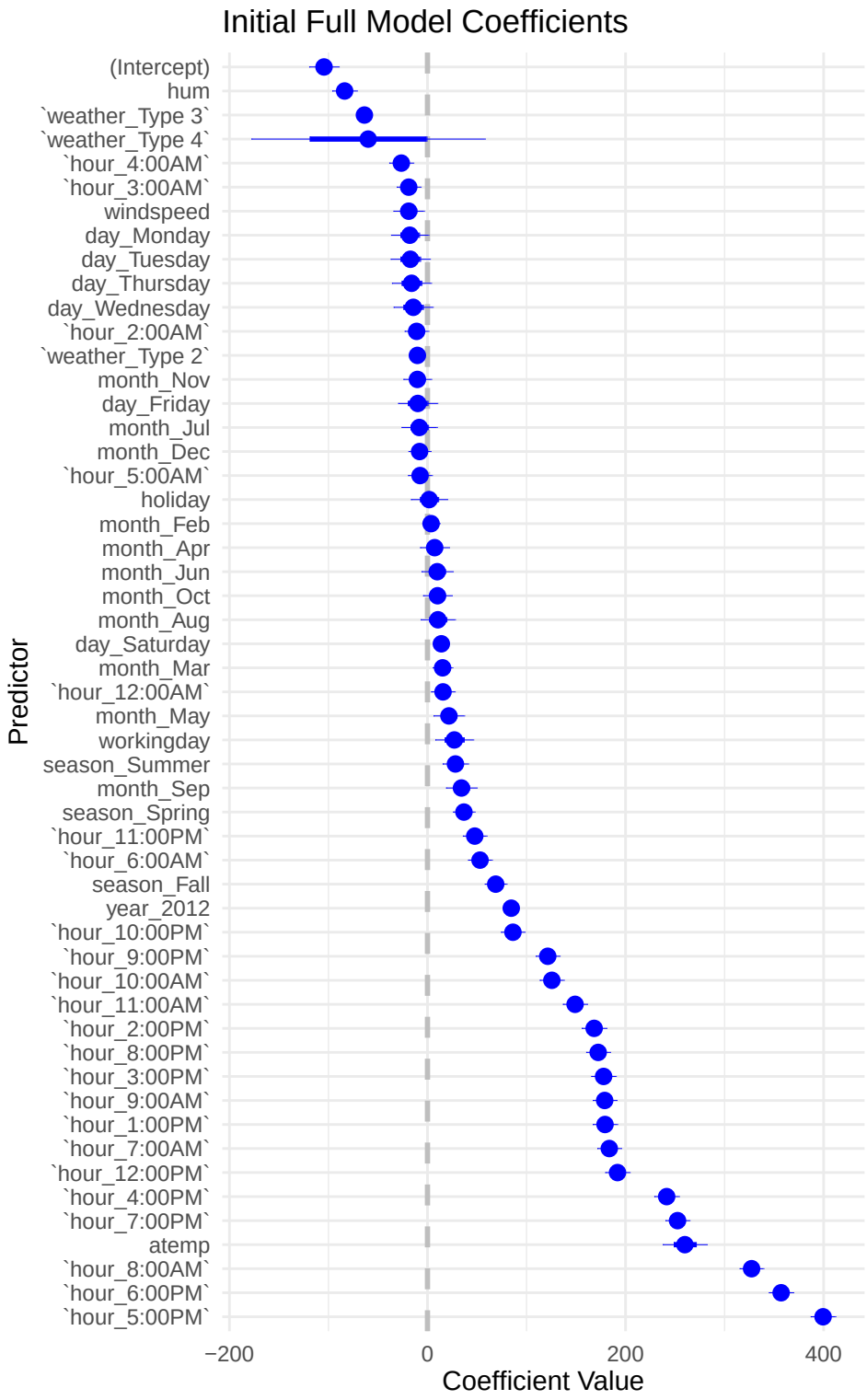

Figure 16: Zero in Confidence Intervals - Insignificant Predictor

```
##
##  Asymptotic one-sample Kolmogorov-Smirnov test
##
## data:  full_model$residuals
## D = 0.51982, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

```
## [1] "H0 rejected: Residuals are NOT normally distributed"
```

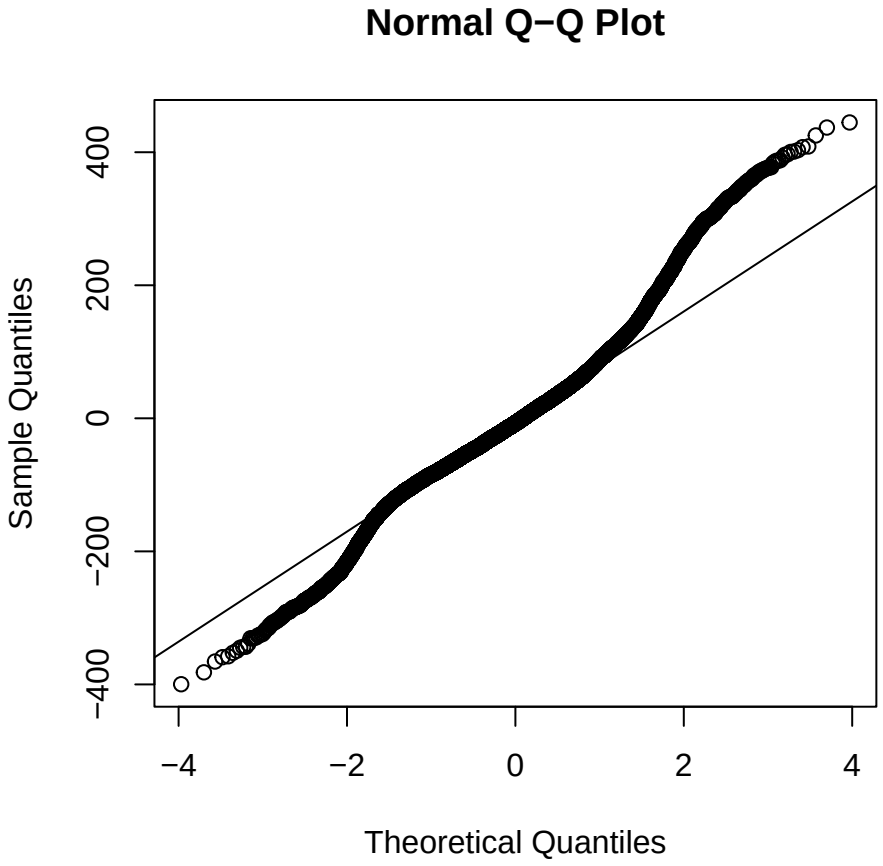

Figure 17: Normality of Residuals Assumption - Violated

```
## [1] "H0 rejected: Error variance spread INCONSTANTLY (Heteroscedasticity)"
```

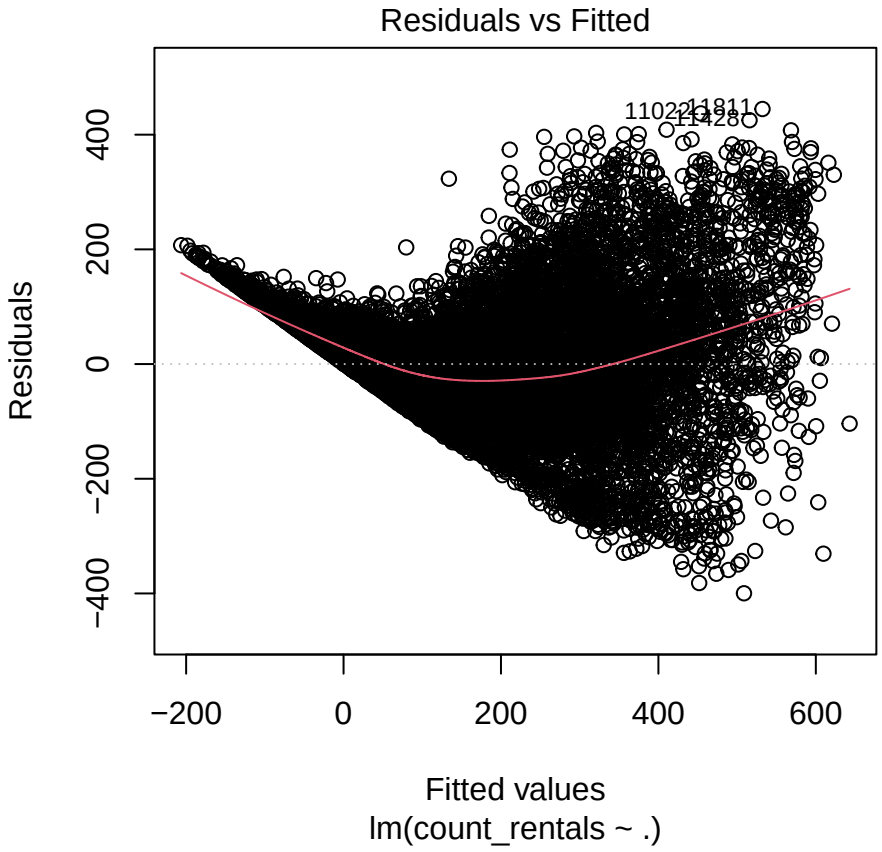

Figure 18: Homoscedasticity and Linearity Assumptions - Violated

```
## [1] Below: Highest Variance Inflation Factors of Full Model
```

```
##                     VIF
## workingday    26.74941
## day_Wednesday 16.17893
## day_Tuesday   16.12691
## day_Thursday  15.89446
## day_Friday    15.69071
## day_Monday    14.79910
## season_Summer 10.61741
## month_Jul      8.25986
```

```
## [1] ---
```

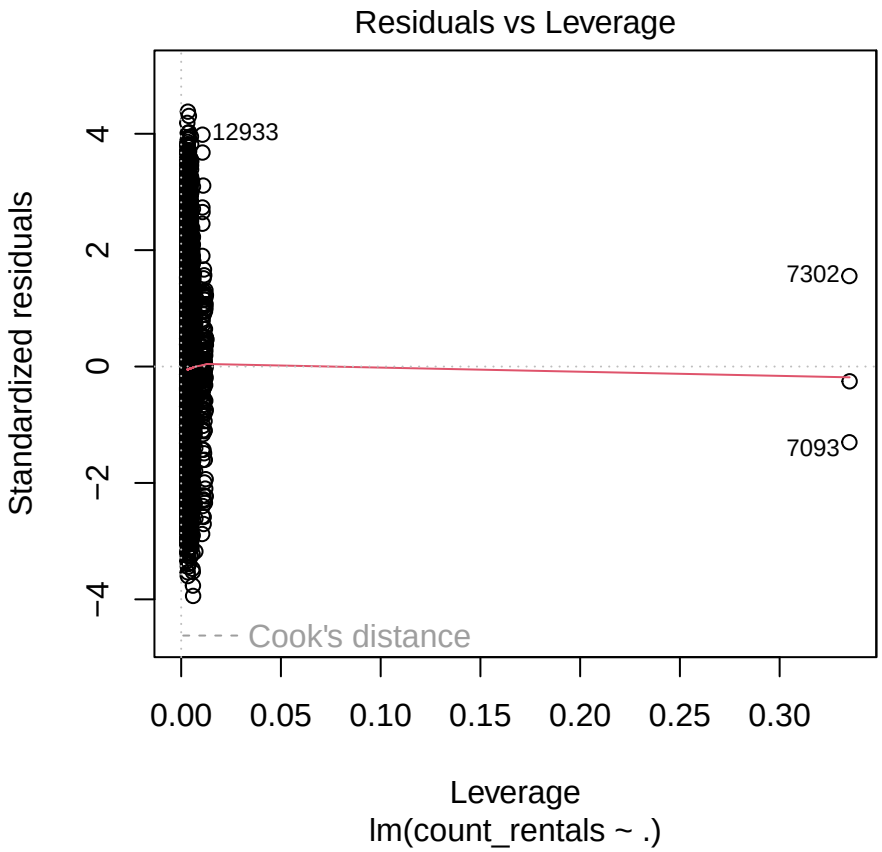

Figure 19: Influential Observations

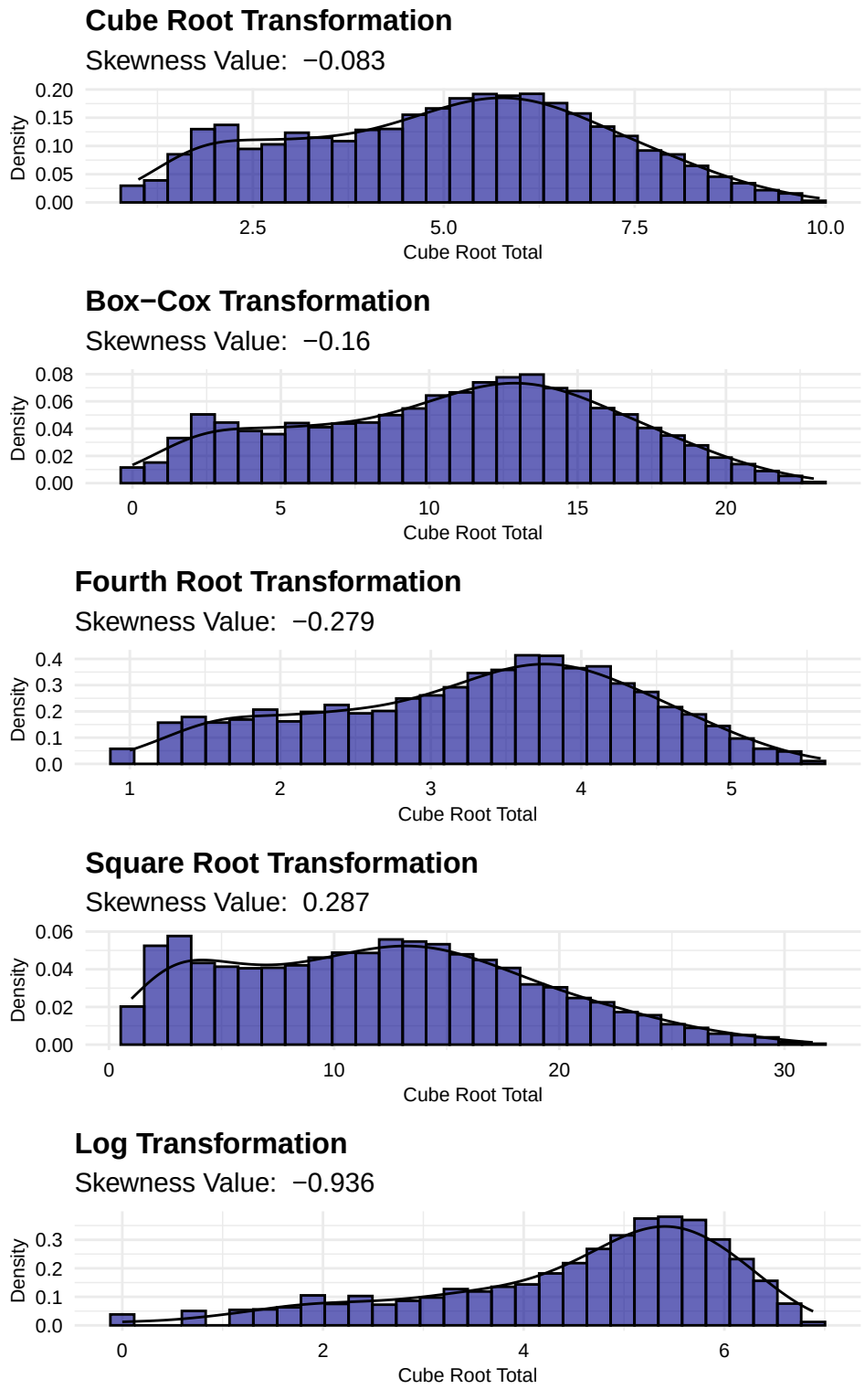

Figure 20: Cube Root Transformation Closest Skewness Value to Zero

Final Model Equation

$$\texttt{cube\_root\_total} = \beta_0 + \beta_1\texttt{holiday} + \beta_2\texttt{atemp} + \beta_3\texttt{hum} + \beta_4\texttt{season\_Spring} + \beta_5\texttt{season\_Fall} + \beta_6\texttt{year\_2012} + \beta_7\texttt{month\_May} + \beta_8\texttt{month\_Aug} + \beta_9\texttt{month\_Sep} + \beta_{10}\texttt{day\_Friday} + \beta_{11}\texttt{day\_Saturday} + \beta_{12}\texttt{hour\_1:00PM} + \beta_{13}\texttt{hour\_10:00AM} + \beta_{14}\texttt{hour\_10:00PM} + \beta_{15}\texttt{hour\_11:00AM} + \beta_{16}\texttt{hour\_11:00PM} + \beta_{17}\texttt{hour\_12:00AM} + \beta_{18}\texttt{hour\_12:00PM} + \beta_{19}\texttt{hour\_2:00AM} + \beta_{20}\texttt{hour\_2:00PM} + \beta_{21}\texttt{hour\_3:00AM} + \beta_{22}\texttt{hour\_3:00PM} + \beta_{23}\texttt{hour\_4:00AM} + \beta_{24}\texttt{hour\_4:00PM} + \beta_{25}\texttt{hour\_5:00AM} + \beta_{26}\texttt{hour\_5:00PM} + \beta_{27}\texttt{hour\_6:00AM} + \beta_{28}\texttt{hour\_6:00PM} + \beta_{29}\texttt{hour\_7:00AM} + \beta_{30}\texttt{hour\_7:00PM} + \beta_{31}\texttt{hour\_8:00AM} + \beta_{32}\texttt{hour\_8:00PM} + \beta_{33}\texttt{hour\_9:00AM} + \beta_{34}\texttt{hour\_9:00PM} + \beta_{35}\texttt{weather\_Type\_2} + \beta_{36}\texttt{weather\_Type\_3} + \epsilon$$

```
## [1] Below: Variance Inflation Factors of Final Model
```

```
##          holiday            atemp              hum    season_Spring
##         1.010416         1.416244         1.790403         1.694333
##      season_Fall        year_2012        month_May        month_Aug
##         1.253539         1.015510         1.428385         1.333166
##        month_Sep       day_Friday     day_Saturday    `hour_1:00PM`
##         1.177779         1.031077         1.034515         2.055870
##   `hour_10:00AM`   `hour_10:00PM`   `hour_11:00AM`   `hour_11:00PM`
##         1.945911         1.948927         1.970123         1.941967
##   `hour_12:00AM`   `hour_12:00PM`    `hour_2:00AM`    `hour_2:00PM`
##         1.955580         2.004225         1.929255         2.038647
##    `hour_3:00AM`    `hour_3:00PM`    `hour_4:00AM`    `hour_4:00PM`
##         1.910124         2.075169         1.890139         2.043925
##    `hour_5:00AM`    `hour_5:00PM`    `hour_6:00AM`    `hour_6:00PM`
##         1.926370         2.010121         1.935278         2.006346
##    `hour_7:00AM`    `hour_7:00PM`    `hour_8:00AM`    `hour_8:00PM`
##         1.923501         1.981749         1.939721         1.963228
##    `hour_9:00AM`    `hour_9:00PM` `weather_Type 2` `weather_Type 3`
##         1.923534         1.935056         1.179746         1.277333
```

```
## [1] ---
```

```
##
##  Asymptotic one-sample Kolmogorov-Smirnov test
##
## data:  stepwise_model$residuals
## D = 0.055945, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

```
##
##  studentized Breusch-Pagan test
##
## data:  stepwise_model
## BP = 3813.8, df = 36, p-value < 2.2e-16
```

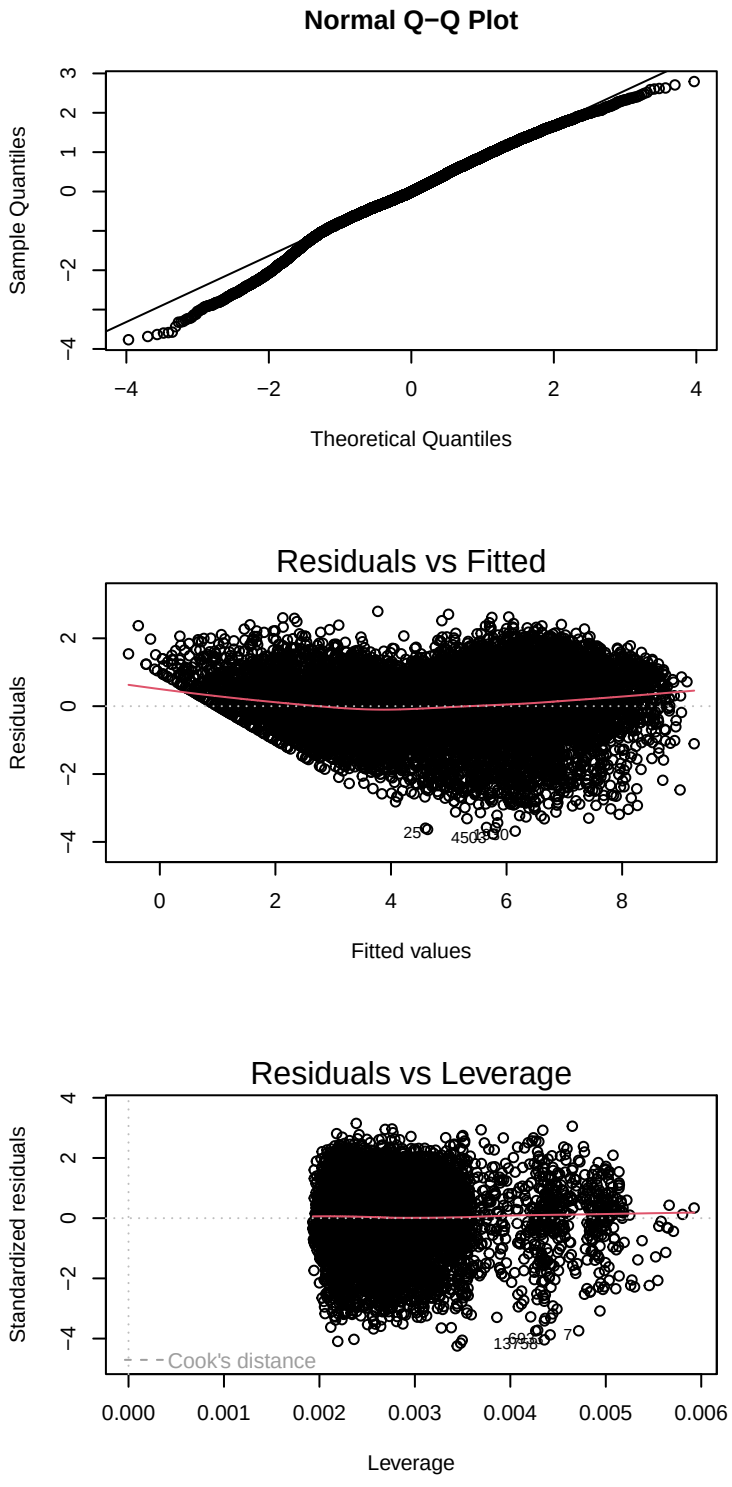

Figure 21: Assumption Corrections and High Leverage Removal

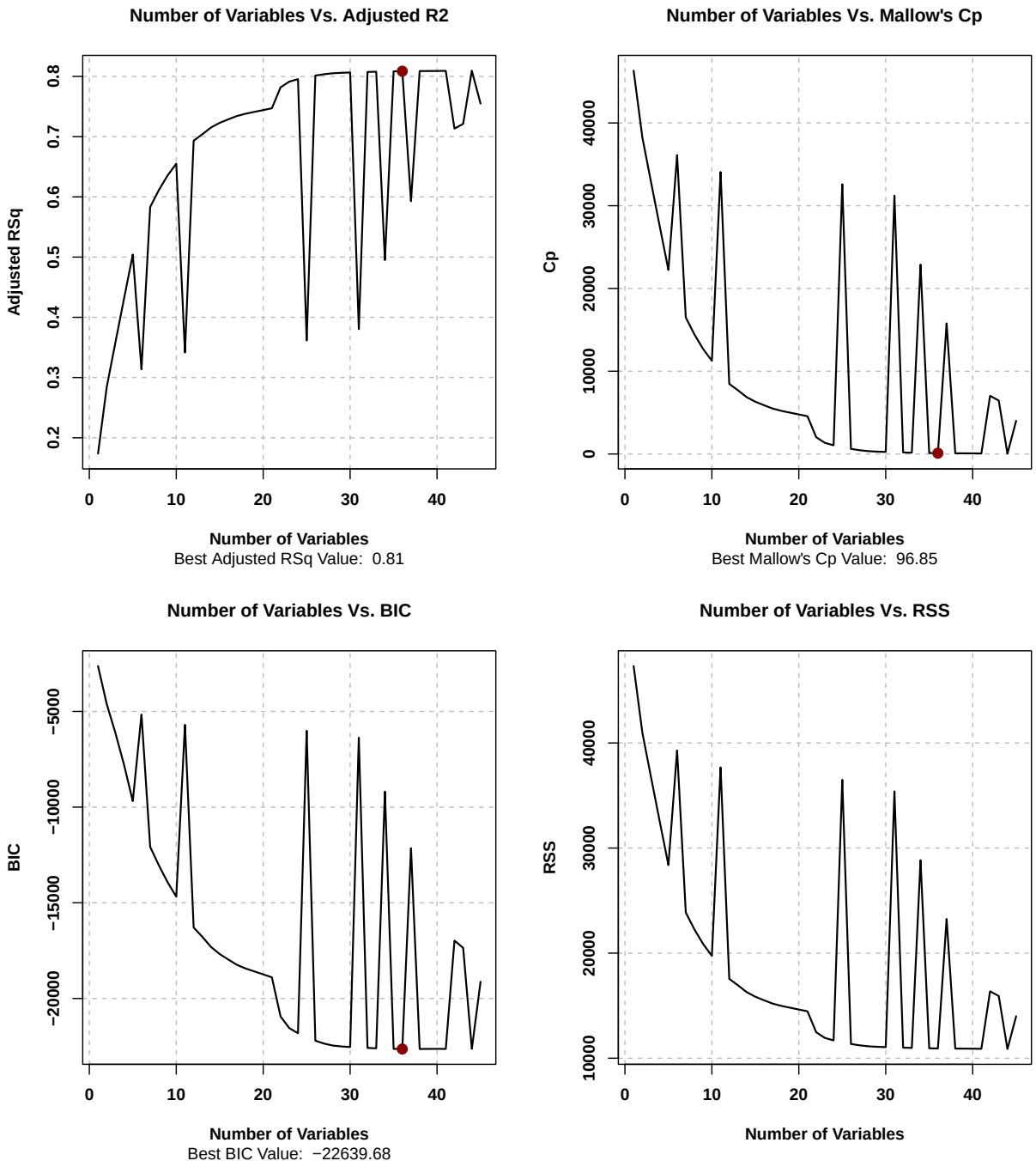

Figure 22: Stepwise Selection Model Evaluations(BIC/RSquared/Mallows Cp/RSS)

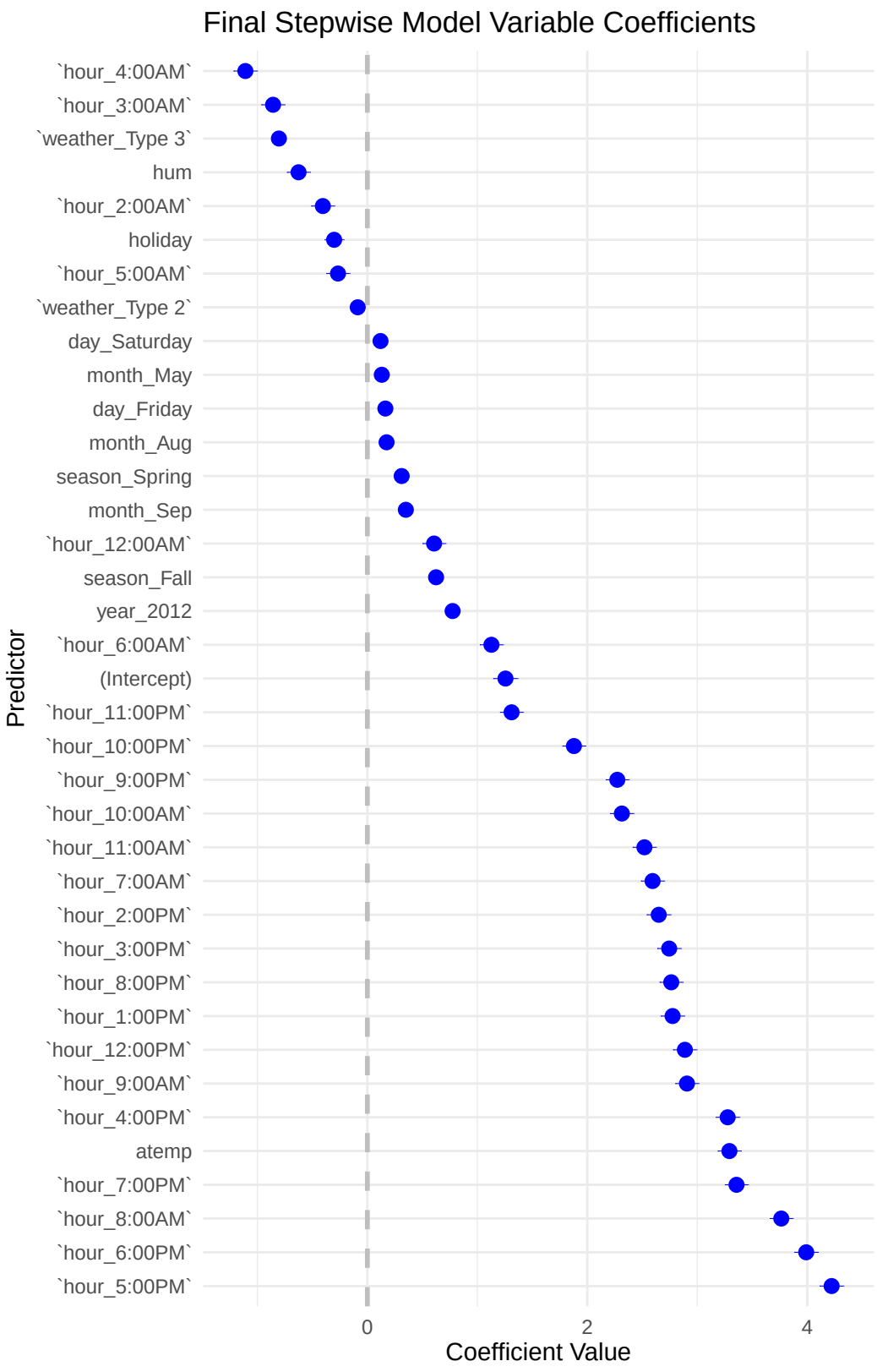

Figure 23: Predictors for Final Stepwise Model

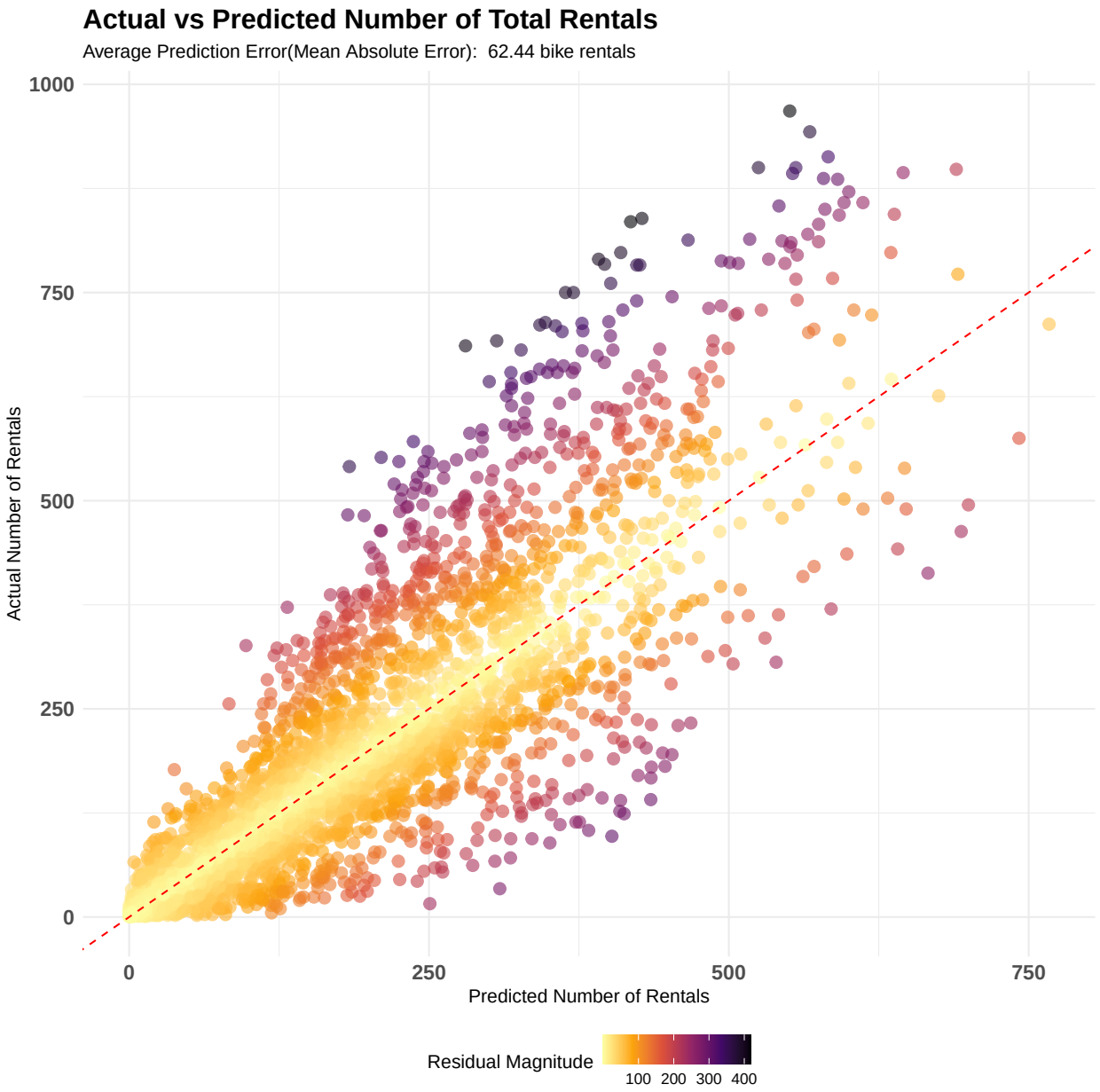

Figure 24: Predictions Using the Test Dataset